

Towards a Transportable Causal Network Model Based on Observational Healthcare Data

Alice Bernasconi^{1,2}, Alessio Zanga^{1,3}, Peter J.F. Lucas⁴, Marco Scutari⁵ and Fabio Stella¹

¹*Models and Algorithms for Data and Text Mining Laboratory (MADLab), University of Milano - Bicocca, Milan, Italy*

²*Evaluative Epidemiology Unit, Department of Epidemiology and Data Science, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy*

³*Data Science and Advanced Analytics, F. Hoffmann - La Roche Ltd, Basel, Switzerland*

⁴*University of Twente, Enschede, The Netherlands*

⁵*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland*

Abstract

Over the last decades, many prognostic models based on artificial intelligence techniques have been used to provide detailed predictions in healthcare. Unfortunately, the real-world observational data used to train and validate these models are almost always affected by biases that can strongly impact the outcomes validity: two examples are values missing not-at-random and selection bias. Addressing them is a key element in achieving transportability and in studying the causal relationships that are critical in clinical decision making, going beyond simpler statistical approaches based on probabilistic association.

In this context, we propose a novel approach that combines selection diagrams, missingness graphs, causal discovery and prior knowledge into a single graphical model to estimate the cardiovascular risk of adolescent and young females who survived breast cancer. We learn this model from data comprising two different cohorts of patients. The resulting causal network model is validated by expert clinicians in terms of risk assessment, accuracy and explainability, and provides a prognostic model that outperforms competing machine learning methods.

Keywords

Causal discovery, Causal networks, Transportability, Missing values, Selection bias

1. Introduction

An important application of artificial intelligence in healthcare is predicting a disease trajectory conditional on the patient's history and a projected treatment strategy. In this context, we aim to develop a model that generalizes well: it must provide accurate predictions not only on the *study cohort* (the population from which the data have been collected) but also on the *target cohort*, the general population that it is designed for. The model's *validity*, accuracy and usefulness in clinical practice are then of paramount relevance.

In order to establish validity in a rigorous causal framework, we must clearly identify the scope of the study the data come from, as well as the applicability and limitations of the theory

HC@AIxIA 2023: 2nd AIxIA Workshop on Artificial Intelligence For Healthcare

✉ a.bernasconi6@campus.unimib.it (A. Bernasconi)

🆔 0000-0001-8522-6882 (A. Bernasconi); 0000-0003-4423-2121 (A. Zanga); 0000-0001-5454-2428 (P.J.F. Lucas); 0000-0002-2151-7266 (M. Scutari); 0000-0002-1394-0507 (F. Stella)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

we are studying. However, many other issues can impact model validity. Firstly, study cohorts differ from target cohorts in both randomized controlled trials and clinical observational studies due to their inclusion and exclusion criteria. Secondly, *missingness patterns* may introduce bias in the model when the data are *missing not at random* (MNAR). Thirdly, information can be fragmented across different data sets (*data sparsity*) and experts (*domain knowledge*). Therefore, it is important to understand the data selection mechanism and the merging process to ensure the model validity and generalisability [1].

Our main contributions in this context are:

- Developing the first causal network model for estimating the risk of cardiovascular diseases in *adolescent and young adults* that have been treated and survived *breast cancer*.
- Developing and adapting a methodological approach to deal with data affected by selection bias and suffering by MNAR.
- A thoughtful review of the model we developed and of its implications by domain experts (that is, expert physicians).

The rest of the paper is organized as follows. We start by introducing our research question and the topic of the study: *cardio-oncology*, the subfield of cardiology that aims at significantly reducing cardiovascular morbidity and mortality and at improving the quality of life in cancer survivors, here young patients (Section 2). We complete this introduction by providing an overview of the related work available in the literature (Section 3). We then describe the data (Section 4.1), the domain knowledge (Section 4.2) and the causal networks methodology (Section 5) that we use to design and develop our model. Finally, we describe our findings (Section 6), we contrast them with the available clinical and epidemiological knowledge and we assess model performance against that of other commonly-used machine learning approaches. We complete the paper by summarising our conclusions and presenting future work (Section 7).

2. Young Adult Breast Cancer Survivors

The increasing number and life expectancy of cancer survivors in the last decades has highlighted the acute and long-term cardio-toxic effects of different cancer therapies. The most common cancer among *adolescent and young-adult* females (AYAs; 15–39 years at the first cancer diagnosis) is breast cancer (BC), which is a rare disease with unique genetic and biological features in this age class. AYA female patients are more likely to be genetically susceptible to more aggressive form of BC than older females, and their diagnosis is often delayed because of the lack of BC screening policies for their age group. While lower than in older females with BC, AYAs BC survival rates are high and continuously increasing, thus making such patients likely to be long-term survivors [2]. Moreover, BC is biologically more aggressive in AYAs than in older women with BC, and requires more aggressive combined neoadjuvant (pre-surgery) and adjuvant (post-surgery) treatments [3].

While cardiovascular diseases (CVDs) are well characterized in some groups of cancer survivors, only few studies describe CVDs in AYAs because of their rarity in young patients. Therefore, further evidence is needed to understand and prevent CVDs in AYAs with BC for

helping clinicians to plan personalized and effective follow-up guidelines for these patients. Thus, with this work we are interested in answering to the question: “*To what extent is it possible to predict and explain individual susceptibility to cardiotoxicity in AYA with BC?*”

3. Related works

Machine learning (ML) techniques have achieved remarkable results in predicting CVDs in cancer patients [4]. Unfortunately, they are still not part of the clinical practice: cardio-oncologists rely on older, less accurate cardiovascular risk stratification tools such as the Framingham score [5]. The reasons for their reluctance to use ML methods are as follows:

- **Data Availability.** The best-performing ML models such as XGBoost and neural networks rely on biomarkers, laboratory tests, electrocardiograms, echocardiograms, computerized tomography, and cardiac magnetic resonance imaging data [6]. However, cardio-oncologists rarely have immediate and complete access to such data in their everyday practice.
- **Computational Burden and Data Scarcity.** ML models are known to be computationally intensive and to require large amounts of data to achieve optimal performance. This limits their use in most of healthcare systems, even in rich countries, because healthcare data suffer from quality issues and are typically scarce. They are also often biased and affected by missing-not-at-random patterns [7] in ways ML models do not account for.
- **High Skills, Knowledge and Expertise.** Current state-of-the-art ML methods require specialized skills, knowledge and expertise to train, validate and deploy especially when combining different types of data, which is almost always the case in healthcare [8].
- **Lack of Interpretability.** The current state-of-the-art ML models are still difficult to interpret despite recent progress from Explainable AI in addressing this issue [9].
- **Lack of generalizability.** ML models have been applied successfully to childhood cancer survivors [10], where using a limited number of variables, a relatively simple model, and an easy-to-understand user-interface ensured that they were incorporated into clinical practice. However, cancer is a collection of very complex and heterogeneous diseases, making ML models unlikely to transfer successfully to other cancer survivors cohorts.
- **Correlation vs Causation.** ML models are typically developed using observational data, and in general they can achieve excellent predictive accuracy by leveraging the *association* between the response and the observed variables. However, clinicians operate in a *causal* framework [11] where they need to evaluate what the outcome could be when prescribing a particular treatment.

4. Materials

4.1. Data

This project makes use of data coming from two separate cohorts:

- **The population-based cohort (PBC):** a retrospective cohort of about 1,500 AYAs BC patients who completed cancer treatment, that is, that survived at least 1 year after the cancer diagnosis. In this cohort, BC cases have been identified in population-based cancer registries and the information from each BC patient has been linked to several administrative data sets (hospital discharge records, outpatients and drug flows) for CVD follow-ups [12].
- **The clinical-based cohort (CBC):** a retrospective single-institution clinical cohort consisting of 340 additional BC patients, with additional detailed information on cancer prognostic factors but with no CVD follow-up information.

Around 3% of AYAs BC patients from the PBC had at least one CVD event during the follow-up (mean follow-up time = 5 years). However, the Framingham score for women with the same characteristics as those in the PBC ranges between 0 and 9, which translates to a predicted risk of 0% to 1% of developing a hard coronary heart disease (such as myocardial infarction) or of dying for a CVD event within 10 years [5]. Exploring the PBC further, we found that the most frequent CVDs were due to chemotherapy-induced cardiac damages (arrhythmia and heart failure, in around 1% of patients) and to ischemic heart diseases that may be related to hormone therapy (in around 2% of patients). Moreover, 15% of patients developed at least one major cardiac risk factor (dyslipidemia, diabetes and hypertension) after cancer treatment, thus resulting in an increase in the corresponding Framingham CVD risk.

The CBC consists of patients treated by a single institution (Fondazione IRCCS Istituto Nazionale dei Tumori di Milano) and is biased due to the patient selection mechanism in ways that are apparent from the baseline and treatment variables distribution. Around 8% of AYAs BC patients from the CBC have at least one major cardiac risk factor before treatment, compared to only 5% in the PBC. Moreover, there is an higher proportion of patients receiving neoadjuvant treatments (30% in the CBC vs 23% in the PBC), which are more likely to be prescribed when dealing with later tumor stages at diagnosis (with lymph nodes involvement, metastasis and high dimension tumors). These facts reflect a more severe case-mix of both baseline characteristics of the patients and of tumor aggressiveness in the CBC compared to the PBC, which is expected since the Institute is a referral expertise cancer center.

Furthermore, the PBC and the CBC were collected for different purposes, resulting in different *missingness mechanisms*. In the CBC, the purpose is patient care, so factors that may influence tumor prognosis (such as tumor grading, staging, etc.) are recorded in greater detail; whereas cancer registry data, which form the basis of the PBC, are collected for administrative, epidemiological monitoring and public health purposes. More details about the distribution of missing values by cohort type is reported in Table S1 in the Supplementary Material. their percentage of completeness in the cohort of origin.

4.2. Clinical Knowledge on Cardiovascular Diseases

In this section we elicit the clinical knowledge on *cardio-oncology*. To make it easier for the reader to understand how the domain knowledge has been translated from natural language to the structure of the causal network model, we report in square brackets the label of the corresponding nodes.

The description of the selection mechanism behind the selection node [cohort] has already been described in Section 4.1.

To decide which treatment to give to a young BC patient, clinicians rely on the most well-known prognostic factors for 5-year cancer survival [death_in_5y] [13]: age (below or above 35 years) [age35], tumor grade [grade], tumor histology [histology], ki67+ status [ki67], molecular subtype [receptors], vascular invasion [vascular], lymph nodes involvement [pN] and tumor dimension [pT]. For example, *triple negative BC* is a specific molecular subtype of BC that is more common in AYAs than in other age groups (prevalence 15%–20% [14]) and whose treatment options are limited to chemotherapy and/or radiotherapy because target therapies are ineffective.

The risk of CVDs after both neoadjuvant [_neo] and adjuvant [_adju] cancer treatments is well known. For instance, anthracycline is frequently prescribed as a chemotherapy regimen [chemo_], either alone or in combinations with other chemotherapeutic drugs. Its cardiotoxic effects are well documented and largely attributable to the generation of free radicals: it ultimately results in left-ventricular dysfunction, arrhythmias [cardiotoxicity] and, in turn, heart failure in the most severe cases [15] [cvds]. Even though radiotherapy [radio_] reduces the risk of cancer recurrence and death in BC patients, the heart may be incidentally exposed to ionizing radiation when the primary cancer is located in the left breast, with in turn increases the risk of heart diseases [cvds] [16]. In addition, target therapy [target_] has been shown to induce acute cardiac toxicity [cardiotoxicity], especially when administered as an adjuvant treatment (say, trastuzumab for young BC patients [17]), due to possible changes in oxidative stress that reflect mainly in left-ventricular-ejection fraction reduction. However, it is generally believed that the trastuzumab-induced cardiotoxic effects are reversible and that they do not impact long-term CVD risk [18].

Moreover, 5 years of tamoxifen [hormons_], with or without ovarian suppression/ablation, is considered the standard hormone therapy in young women with hormone-receptor positive disease, which represent the vast majority in this age group. However, it has frequent chronic late effects like type-2 diabetes [t2db], hypertension [hypertension] and dyslipidemia [dyspidemia] which are all known to be major risk factors for long-term ischemic heart diseases [ischemic_heart_disease] [19].

5. Methods

Assessing validity is a crucial aspect of causal inference. Specifically, *internal validity* refers to *generalizing* the evidence from a study sample to the underlying *study population*; *external validity* refers to *transporting* the evidence to a different *target population*. The interest in the transportability of inference has increased exponentially in the last few years [20, 21] thanks to methodological breakthroughs in handling distribution shifts and selection bias. In this

work, we leverage both *missingness graphs* [22, 23, 24] and *selection diagrams* [25] to learn a causal Bayesian network by combining observational data together with prior knowledge in the presence of selection bias and missingness bias.

5.1. Causal Models for Missing Values and Selection Bias

Model-based approaches rely on the formal specification of a *model* representing the interactions we are interested in. In the case of causal inference, *causal graphs* are the de-facto standard for encoding causal relationships.

Definition 1 (Causal Graph). A causal graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ [26] is a directed acyclic graph where for each directed edge $(X, Y) \in \mathbf{E}$, X is a direct cause of Y and Y is a direct effect of X . The vertex set \mathbf{V} can be split into two disjoint subsets $\mathbf{V} = \mathbf{O} \cup \mathbf{U}$, where \mathbf{O} is the set of the fully observed variables (with no missing values) and \mathbf{U} is the set of fully unobserved variables, also known as latent variables.

A causal graph allows researches to (i) decide if a consistent estimator exists for a casual effect and (ii) derive that estimator directly from the graph. Being able to correctly estimate causal effects is extremely important for policy making. For instance, it allows clinicians to assess the impact of a given drug on a disease from observational studies when randomized controlled trials are not available for a particular sub-population [27]. In more general terms, let \mathcal{G} be a causal graph, X a treatment and Y an outcome. A consistent estimator for the causal effect of X on Y is given by the *do-operator*:

$$P(Y = y | do(X = x)) = \sum_{\mathbf{z}} P(Y = y | X = x, \mathbf{Z} = \mathbf{z}) P(\mathbf{Z} = \mathbf{z}) \quad (1)$$

if there exists a set \mathbf{Z} that satisfies the *back-door criterion* for \mathcal{G} [28]. If no consistent estimator exists, *confounding bias* makes it impossible to estimate the causal effect correctly. Unbiased estimators of causal effects are theoretically possible even without a causal graph, but their assumptions are rarely satisfied in practical applications [29]. Therefore, causal graphs remain the tool of choice to achieve both explainability and consistency.

The *missingness mechanism* also plays an important role in causal modeling. Common pre-processing techniques that deal with missing values such as sample deletion and missing imputation are often ineffective or even detrimental to causal effect estimation. For instance, [30, 31] have shown how missing-data handling has a significant impact on the clinical conclusions that can be drawn from experimental results. Therefore, modeling the reason why a given value is missing is crucial, especially for *MNAR* data.

The framework for modeling missingness mechanisms is detailed in Rubin’s foundational work [32]. More recently, *missingness graphs* [33, 22] have been proposed to reconcile Rubin’s framework with causal graphs by directly including the missingness indicators in an extended graph structure.

Definition 2 (Missingness Graph). A missingness graph $\mathcal{M} = (\mathbf{V}, \mathbf{E})$ [22] is a causal graph whose vertex set \mathbf{V} is partitioned into five disjoint subsets $\mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \cup \mathbf{S} \cup \mathbf{R}$, where: \mathbf{M} is the

set of the partially observed variables, that is, the variables with at least one missing value; \mathbf{S} is the set of the proxy variables, that is, the variables that are actually observed; \mathbf{R} is the set of the missingness indicators.

Since missingness graphs are extended causal graphs, *d-separation* [34] implies (conditional) independence. Briefly, a set of variables \mathbf{Z} d-separates X from Y , denoted by $X \perp\!\!\!\perp Y \mid \mathbf{Z}$, if it *blocks* every path (the combination of all edges and nodes that connect two selected nodes of interest) between X and Y . A path is blocked by \mathbf{Z} if and only if it contains: a fork $A \leftarrow B \rightarrow C$ or a chain $A \rightarrow B \rightarrow C$ so that B is in \mathbf{Z} , or, a collider $A \rightarrow B \leftarrow C$ so that B , or any descendant of it, is not in \mathbf{Z} . In this framework, there exists a one-to-one correspondence between the *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *MNAR* patterns and the independence statements implied by the missingness graph:

- MCAR implies $\mathbf{O} \cup \mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R}$: missingness is random and independent from the fully observed variables \mathbf{O} and the partially observed variables \mathbf{M} ;
- MAR implies $\mathbf{U} \cup \mathbf{M} \perp\!\!\!\perp \mathbf{R} \mid \mathbf{O}$: missingness is random only conditionally on the fully observed variables \mathbf{O} ;
- MNAR if neither MCAR nor MAR.

This makes it possible to verify whether a consistent estimator for the joint probability $P(\mathbf{X})$ exists in case of MNAR. When the conditions in [33] hold, $P(\mathbf{X})$ is *recoverable* and a consistent estimator is given by

$$P(\mathbf{X}) = \frac{P(\mathbf{R}_{\mathbf{X}} = \mathbf{0}, \mathbf{X})}{\prod_{X \in \mathbf{X}} P(R_X = 0 \mid \Pi_{R_X}, \mathbf{R}_{\Pi_{R_X}} = \mathbf{0})}, \quad (2)$$

where R_X is the missingness indicator for variable X , Π_{R_X} is the parent set of R_X and $\mathbf{R}_{\mathbf{X}}$ is the union of the R_X for all the variables in \mathbf{X} .

When the data are a collation of multiple data sets, their different *selection criteria* may induce discrepancies in the distribution of some of the collected variables. In such cases, pooling the data together without modeling the *context* from which observations come from could induce inconsistent estimates [35]. *Selection diagrams* are introduced as an extension of causal graphs for this purpose. We report the definition for completeness.

Definition 3 (Selection Diagram). Let Π and Π^* be two different populations with a common underlying causal graph \mathcal{G} . A selection diagram \mathcal{S} extends the causal graph \mathcal{G} so that:

- $\mathbf{E}_{\mathcal{G}} \subset \mathbf{E}_{\mathcal{S}}$, that is, \mathcal{G} is a sub-graph of \mathcal{S} ;
- $\exists \mathbf{S} \neq \emptyset, \mathbf{S} \subset \mathbf{V}_{\mathcal{S}} \wedge \mathbf{S} \not\subset \mathbf{V}_{\mathcal{G}}$ where $(S_i \rightarrow V_i) \in \mathbf{E}_{\mathcal{S}}$ iff $f_i \neq f_i^*$, that is, the variables \mathbf{S} point to the variables $\mathbf{V}_{\mathcal{G}}$ that differ in their value assignment f in Π and Π^* .

The variables \mathbf{S} are usually called *selection variables* and allow us to pool together all the observations while keeping track of their provenance. In related work on selection bias, these variables are called *context variables* and identify the context in which $\mathbf{V}_{\mathcal{G}}$ have been collected. We refer the reader to [36] for an extended discussion of the differences between selection variables and context variables. In our setting we can use these terms interchangeably. Intuitively,

selection bias represents the difference between a consistent estimate for Π and a consistent estimate for Π^* : it limits our ability to *transport* the inference made on a model for a study population Π to another target population Π^* due to the differences in their selection criteria. As for d-separation, when a selection diagram \mathcal{S} is modeled, it is possible to identify a consistent estimator for the given set of selection variables \mathbf{S} using the *g-transportability* [37] criterion.

Once a causal graph \mathcal{G} is designed, we need to connect it to the data \mathcal{D} . To do so, we rely on *Causal networks* (CNs) [38], an extension of the well known *Bayesian networks* (BNs) [39], where the underlying graph is a causal graph.

Definition 4 (Causal Network). Let be \mathcal{G} a causal graph and let $P(\mathbf{X})$ be a global probability distribution with parameters Θ . A causal network $\mathcal{C} = (\mathcal{G}, \Theta)$ is a causal model where each variable of \mathbf{X} is a vertex of \mathcal{G} and $P(\mathbf{X})$ factorizes into local probability distributions following:

$$P(\mathbf{X}) = \prod_{X \in \mathbf{X}} P(X | \Pi_X) \quad (3)$$

where Π_X is the parent set of the variable X .

If we combine a missingness graph and a selection diagram into a single causal graph, we can recover from confounding bias, inconsistent estimators due to missing values and domain discrepancies with a single causal network.

5.2. Causal Discovery with Missing Values and Prior Knowledge

Manually constructing the *true* causal graph is virtually impossible in real-world applications where there is little to no control over the experimental setting. For instance, there could be unobserved variables affecting the data generating mechanism or unknown interactions between the observed variables. *Causal discovery* [40, 41] focuses on recovering the causal graph from collected data and prior knowledge. The existing literature typically assumes that there are no missing values or that the missingness mechanism is either MCAR or MAR [42]; extensions to MNAR have only appeared in recent years [23, 24].

In this contribution, we take into account the impact of the missingness mechanism using the *Structural Expectation-Maximization* (SEM) algorithm [43, 44, 45] and clinical prior knowledge. At its core, SEM repeats two steps until convergence is reached:

- Expectation (E): impute the missing values $\widehat{\mathcal{D}}_i$ from $\widehat{\mathcal{C}}_i$ by estimating $\widehat{\Theta}_i$ using $\widehat{\mathcal{G}}_i$.
- Maximization (M): find $\widehat{\mathcal{G}}_{i+1}$ that maximises a given score function for $\widehat{\mathcal{D}}_i$.

At the end of the i th iteration, SEM produces an estimated causal network $\widehat{\mathcal{C}}_{i+1}$ that will be the starting point of the $(i + 1)$ th iteration. In the case of MNAR, it is essential to model the missingness mechanism *a priori* in order to avoid spurious associations induced by the missingness pattern. We achieve that by providing an initial \mathcal{G}_0 that encodes a *partially-specified* missingness graph from prior knowledge and by fixing its arcs in place throughout the causal discovery, effectively assuming that they correctly specify the missingness mechanism. Hence, the SEM algorithm only *extends* the initial graph, looking for a super-graph of \mathcal{G}_0 that better fits the collected data \mathcal{D} .

5.3. Causal Model Evaluation

Literature validation. We validated the edges that are added by the SEM algorithm to the model using the medical literature to support the results relevance in this domain. For this purpose, we performed a literature review to identify publications that explain what we observe in the data.

Performance metrics. The combined data set comprising the PBC and CBC cohorts was split in a training set and a test set. As mentioned in Section 4.1, the CBC does not include information on the outcome variables, so we included it only in the training set. The ratio between training and test set was 70% (60% PBC + 10% CBC) / 30% (PBC). We considered further splitting the training set to obtain a validation set, but that would reduce the sets sample size to the point of making the analysis unfeasible.

We evaluated the CN we learned from the training set on the test set by estimating the probability of CVDs and the associated Area Under the receiver operating characteristic Curve (AUC). In particular, we contrasted the AUC obtained with \mathcal{G}_0 , with the network learned from SEM (starting from \mathcal{G}_0), and with other standard ML methods. AUC represents the degree or measure of separability: it measures how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting healthy individuals as healthy and individuals affected by CVDs as at risk.

6. Experimental Results

Validation of SEM algorithm results. Figure 1 shows the network learned by SEM: edges elicited from the domain knowledge are in black, while the new ones, that highlight new medical insights, are in **bold blue**. The latter comprise:

- *Edges directed from [cohort] to the three adjuvant treatments [radio_adju], [chemo_adju], [hormons_adju] and to one neoadjuvant treatment [target_neo].* These edges can be explained by the selection mechanism: the treatments nodes act as proxies of case severity [13] which is more likely to be higher in the CBC than in the PBC.
- *Edges directed from the neoadjuvant treatments [chemo_neo] and [target_neo] to adjuvant treatments [chemo_adju], [radio_adju] and [target_adju].* These edges can be explained by surgical details not encoded in the [surgery] node, which only describes whether the surgical approach was radical or conservative. The prioritization of breast preservation in AYAs results in more conservative surgical approaches and, in turn, in a higher probability of incomplete surgical margins. As a result, adjuvant treatments are more frequently used to reduce the risk of both local and distant recurrences [46].
- *Edges directed from target therapy [target_neo] to cardiovascular diseases [cvds].* These edges can be explained by other unstudied effects of oxidative stress and inflammation that induce CVDs but that are not mediated by cardiotoxicity (for instance,

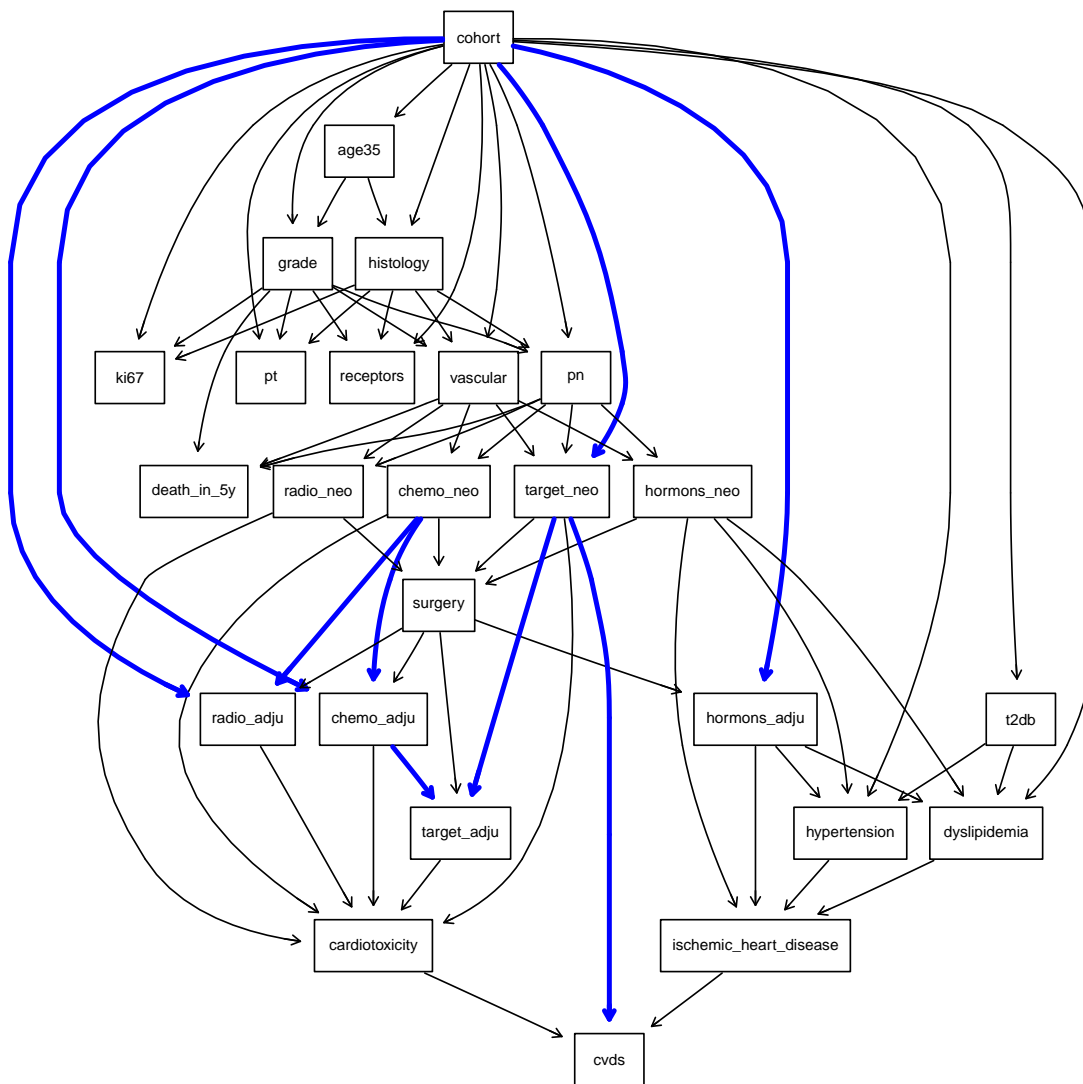


Figure 1: Causal network learned by combining data together with prior knowledge. The edges in black are included in the prior knowledge provided by clinicians, the edges in blue are added by the SEM algorithm.

metabolomic and other unmeasured CV risk factors such as obesity) [47]. Moreover, according to the literature, our understanding of the mechanisms of trastuzumab-mediated cardiotoxicity is still evolving [17]. Therefore, they should be further investigated to better tailor the CVDs follow-up guidelines in young BC survivors.

Model performance. The CN model based on the clinical prior knowledge illustrated in Section 4.2 showed a strong classification performance, with an AUC of about 83% in the test set. The new edges introduced by the SEM algorithm improved performance to about 88%.

It is important to clarify that results obtained from standard ML methods are not directly comparable to those obtained by the proposed approach, for several reasons. Firstly, standard ML methods are not designed to accommodate missing values, relying instead on pre-processing techniques such as sample deletion or imputation. Therefore, only cases coming from the PBC and with complete information on all variables, less than half of the available cases (≈ 800), can be used to train and validate ML models. Secondly, these models do not address the selection bias and lack of external validity. Finally, their results are not causally and clinically interpretable. Against these limitations, we can see in Table 1 that the proposed CN models outperforms all standard ML methods in terms of classification performance.

Table 1

Model performance comparison using Area Under the receiver operating characteristic Curve (AUC).

	Random Forest	XGBoost	AODE	Naive Bayes	CN prior only	CN prior+SEM
AUC	36.2%	65.6%	74.0%	72.7%	82.9%	87.6%

7. Conclusions and Future work

The CN shown in Figure 1 is the first causal network model for estimating the CVD risk in adolescents and young adult females who survived breast cancer. Thanks to the combination of several causal inference methods, including missingness graphs, selection diagrams and causal discovery algorithms, we were able to build a model that combines domain expert knowledge and multiple data sets. The proposed model is able to:

- *Deal with the uncertainty and biases* that are intrinsic to real-world observational data and produce valid results, generalizing to other applications or populations;
- *Combine data coming from different data sources*, which is especially relevant when dealing with small and understudied populations for which data are extremely sparse.
- *Incorporate the domain knowledge* coming from experts, which has been shown to be fundamental in maximizing the model’s performance.
- *Provide interpretable causal recommendations* to identify those patients that are at higher risk of developing a CVDs and what risk factors are involved, to help physicians in clinical decision making and patient follow-up management.

Moreover, the edges added by the causal discovery to the prior-knowledge-based causal network improved its performance and highlighted new arcs that open medical insights particularly valuable, especially for such an understudied population.

According to our experimental results, the proposed CN model outperforms standard supervised learning methods in terms of classification accuracy: this is unexpected since these methods are specifically trained to optimise it. There are two possible explanations for this phenomenon. Firstly, as mentioned in the results section, the data used to train the models are

XGBoost: Extreme Gradient Boosting; AODE: Averaged One-Dependence Estimators.

different: standard ML methods can use only cases coming from the PBC and with complete information on all variables, which means less than half of the cases used to train the CN models. Secondly, the CN models did not learn the model *from scratch* thanks to their ability to incorporate prior domain knowledge; that makes them more efficient at dealing with low sample sizes and missing-not-at-random issues.

Despite its relevance and strengths, this work has some limitations. First of all, we focused on the classification accuracy of a specific variable of interest instead of evaluating the CN as a whole. This allowed us to compare it with commonly-used ML supervised learning approaches using standard metrics such as AUC, arguing that the proposed CN is competitive in this respect. Having shown that, we are now ready to move beyond prediction into more advanced applications of causal inference which are impossible to carry out with traditional ML models. Moreover, while the arcs added by the SEM algorithm have been validated by a clinical expert and a literature review, a further review by a domain expert (that is, a cardio-oncologist) will be crucial to completely validate the final model.

In conclusion, we are working on further extending our model to include unmeasured variables, such as distant metastases; this is necessary to further address the need of a clinically relevant model, to help cardio-oncologists in their everyday practice.

Acknowledgments

Alice Bernasconi is funded by an AIRC 2020 project, grant number 24864, titled “*pPredicting cardiOvascular diSeases iN adolescent and young breast caNcer pAtients (ROSANNA)*”.

Alessio Zanga is funded by F. Hoffmann-La Roche Ltd.

This work was partially supported by the MUR under the grant “Dipartimenti di Eccellenza 2023-2027” of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy.

References

- [1] L. Leung, Validity, reliability, and generalizability in qualitative research, *Journal of Family Medicine and Primary Care* 4 (2015) 324. doi:10.4103/2249-4863.161306.
- [2] A. Trama, L. Botta, R. Foschi, A. Ferrari, C. Stiller, E. Desandes, M. M. Maule, F. Merletti, G. Gatta, Survival of european adolescents and young adults diagnosed with cancer in 2000–07: population-based data from eurocare-5, *The Lancet Oncology* 17 (2016) 896–906. doi:10.1016/S1470-2045(16)00162-5.
- [3] R. Schaffar, C. Bouchardy, P. O. Chappuis, A. Bodmer, S. Benhamou, E. Rapiti, A population-based cohort of young women diagnosed with breast cancer in geneva, switzerland, *PLOS ONE* 14 (2019) e0222136. doi:10.1371/journal.pone.0222136.
- [4] R. Altena, L. Hubbert, N. A. Kiani, Y. Wengström, J. Bergh, E. Hedayati, Evidence-based prediction and prevention of cardiovascular morbidity in adults treated for cancer, *Cardio-Oncology* 7 (2021) 20. doi:10.1186/s40959-021-00105-y.
- [5] W. Law, C. Johnson, M. Rushton, S. Dent, The framingham risk score underestimates

- the risk of cardiovascular events in the her2-positive breast cancer population, *Current Oncology* 24 (2017) 348–353. doi:10.3747/co.24.3684.
- [6] N. Madan, J. Lucas, N. Akhter, P. Collier, F. Cheng, A. Guha, L. Zhang, A. Sharma, A. Hamid, I. Ndiokho, E. Wen, N. C. Garster, M. Scherrer-Crosbie, S.-A. Brown, Artificial intelligence and imaging: Opportunities in cardio-oncology, *American Heart Journal Plus: Cardiology Research and Practice* 15 (2022) 100126. doi:10.1016/j.ahjo.2022.100126.
- [7] H. Chen, D. Ouyang, T. Baykaner, F. Jamal, P. Cheng, J.-W. Rhee, Artificial intelligence applications in cardio-oncology: Leveraging high dimensional cardiovascular data, *Frontiers in Cardiovascular Medicine* 9 (2022). doi:10.3389/fcvm.2022.941148.
- [8] P. Nolan, Artificial intelligence in medicine – is too much transparency a good thing?, *Medico-Legal Journal* (2023) 002581722211412. doi:10.1177/00258172221141243.
- [9] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, *Entropy* 23 (2020) 18. doi:10.3390/e23010018.
- [10] E. J. Chow, Y. Chen, L. C. Kremer, N. E. Breslow, M. M. Hudson, G. T. Armstrong, W. L. Border, E. A. Feijen, D. M. Green, L. R. Meacham, K. A. Meeske, D. A. Mulrooney, K. K. Ness, K. C. Oeffinger, C. A. Sklar, M. Stovall, H. J. van der Pal, R. E. Weathers, L. L. Robison, Y. Yasui, Individual prediction of heart failure among childhood cancer survivors, *Journal of Clinical Oncology* 33 (2015) 394–402. doi:10.1200/JCO.2014.56.1373.
- [11] E. Bareinboim, J. Pearl, Causal inference and the data-fusion problem, *Proceedings of the National Academy of Sciences of the United States of America* 113 (2016) 7345–7352. doi:10.1073/pnas.1510507113.
- [12] A. Bernasconi, G. Barigelletti, A. Tittarelli, L. Botta, G. Gatta, G. Tagliabue, P. Contiero, S. Guzzinati, A. Andreano, G. Manneschi, F. Falcini, M. Castaing, R. A. Filiberti, C. Gasparotti, C. Cirilli, W. Mazzucco, L. Mangone, S. Iacovacci, M. F. Vitale, F. Stracci, S. Piffer, R. Tumino, S. Carone, G. Sampietro, A. Melcarne, P. Ballotari, L. Boschetti, S. Pisani, L. C. D’Oro, F. Cuccaro, A. D’Argenzio, G. D’Orsi, A. C. Fanetti, A. Ardizzzone, G. Candela, F. Savoia, C. Pascucci, M. Castelli, C. Storchi, A. Trama, Adolescent and young adult cancer survivors: Design and characteristics of the first nationwide population-based cohort in Italy, *Journal of Adolescent and Young Adult Oncology* 9 (2020) 586–593. doi:10.1089/jayao.2019.0170.
- [13] F. Cardoso, S. Kyriakides, S. Ohno, F. Penault-Llorca, P. Poortmans, I. Rubio, S. Zackrisson, E. Senkus, Early breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up, *Annals of Oncology* 30 (2019) 1194–1220. doi:10.1093/annonc/mdz173.
- [14] H. J. Kim, S. Kim, R. A. Freedman, A. H. Partridge, The impact of young age at diagnosis (age <40 years) on prognosis varies by breast cancer subtype: A U.S. SEER database analysis, *The Breast* 61 (2022) 77–83. doi:10.1016/j.breast.2021.12.006.
- [15] M. Volkova, R. Russell, Anthracycline cardiotoxicity: Prevalence, pathogenesis and treatment, *Current Cardiology Reviews* 7 (2012) 214–220. doi:10.2174/157340311799960645.
- [16] C. Taylor, A. Kirby, Cardiac side-effects from breast cancer radiotherapy, *Clinical Oncology* 27 (2015) 621–629. doi:10.1016/j.clon.2015.06.007.
- [17] N. Mohan, J. Jiang, M. Dokmanovic, W. J. Wu, Trastuzumab-mediated cardiotoxicity: current understanding, challenges, and frontiers, *Antibody Therapeutics* 1 (2018) 13–17. doi:10.1093/abt/tby003.

- [18] J. An, M. S. Sheikh, Toxicology of trastuzumab: An insight into mechanisms of cardiotoxicity, *Current Cancer Drug Targets* 19 (2019) 400–407. doi:10.2174/1568009618666171129222159.
- [19] A. Christinat, S. D. Lascio, O. Pagani, Hormonal therapies in young breast cancer patients: when, what and for how long?, *Journal of thoracic disease* 5 Suppl 1 (2013) S36–46. doi:10.3978/j.issn.2072-1439.2013.05.25.
- [20] I. Degtiar, S. Rose, A Review of Generalizability and Transportability, *Annual Review of Statistics and Its Application* 10 (2023) 501–524. URL: <https://www.annualreviews.org/doi/10.1146/annurev-statistics-042522-103837>. doi:10.1146/annurev-statistics-042522-103837.
- [21] K. M. Esterling, D. Brady, E. Schwitzgebel, The Necessity of Construct and External Validity for Generalized Causal Claims, Technical Report 18, s.l., 2023. URL: <http://hdl.handle.net/10419/268605>.
- [22] K. Mohan, J. Pearl, Graphical Models for Processing Missing Data, *Journal of the American Statistical Association* 116 (2018) 1023–1037. URL: <https://www.tandfonline.com/doi/full/10.1080/01621459.2021.1874961><http://arxiv.org/abs/1801.03583>. doi:10.1080/01621459.2021.1874961.
- [23] R. Tu, K. Zhang, P. Ackermann, B. C. Bertilson, C. Glymour, H. Kjellström, C. Zhang, Causal Discovery in the Presence of Missing Data (2018). URL: <http://arxiv.org/abs/1807.04010>.
- [24] Y. Liu, A. C. Constantinou, Greedy structure learning from data that contain systematic missing values, *Machine Learning* 111 (2022) 3867–3896. URL: <https://link.springer.com/article/10.1007/s10994-022-06195-8>. doi:10.1007/s10994-022-06195-8/TABLES/9.
- [25] S. Lee, J. Correa, E. Bareinboim, General Transportability – Synthesizing Observations and Experiments from Heterogeneous Domains, *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020) 10210–10217. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/6582>. doi:10.1609/aaai.v34i06.6582.
- [26] J. Pearl, Causal diagrams for empirical research, *Biometrika* 82 (1995) 669–688. doi:10.1093/biomet/82.4.669.
- [27] C. Fernández-Loría, F. Provost, Causal Decision Making and Causal Effect Estimation Are Not the Same... and Why It Matters, the inaugural issue of the *INFORMS Journal of Data Science* (2021). URL: <http://arxiv.org/abs/2104.04103>.
- [28] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed., Cambridge University Press, USA, 2009.
- [29] M. A. Hernán, J. M. Robins, *Causal Inference: What If*, Boca Raton: Chapman & Hall/CRC, 2020.
- [30] M. R. Stavseth, T. Clausen, J. Røislien, How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data, *SAGE Open Medicine* 7 (2019) 205031211882291. URL: <https://doi.org/10.1177/2050312118822912>. doi:10.1177/2050312118822912.
- [31] A. Zanga, A. Bernasconi, P. J. F. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, F. Stella, Causal Discovery with Missing Data in a Multicentric Clinical Study, in: *Proceedings of the 21st International Conference of Artificial Intelligence in Medicine (AIME)*, volume 13897 LNAI, 2023, pp. 40–44. URL: https://link.springer.com/10.1007/978-3-031-34344-5_5. doi:10.1007/978-3-031-34344-5{_}5.

- [32] D. D. Rubin, Inference and missing data, *Biometrika* 63 (1976) 581–592. URL: <https://academic.oup.com/biomet/article-lookup/doi/10.1093/biomet/63.3.581>. doi:10.1093/biomet/63.3.581.
- [33] K. Mohan, J. Pearl, J. Tian, Graphical models for inference with missing data, in: *Advances in Neural Information Processing Systems*, 2013. URL: <https://proceedings.neurips.cc/paper/2013/file/0ff8033cf9437c213ee13937b1c4c455-Paper.pdf>.
- [34] D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, The MIT Press, 2009.
- [35] P. Forré, J. M. Mooij, Causal calculus in the presence of cycles, latent confounders and selection bias, 35th Conference on Uncertainty in Artificial Intelligence, UAI 2019 (2019).
- [36] J. M. Mooij, S. Magliacane, T. Claassen, Joint causal inference from multiple contexts, *Journal of Machine Learning Research* 21 (2020) 1–108.
- [37] E. Bareinboim, J. Pearl, A General Algorithm for Deciding Transportability of Experimental Results, *Journal of Causal Inference* 1 (2013) 107–134. URL: <https://www.degruyter.com/document/doi/10.1515/jci-2012-0004/html>. doi:10.1515/jci-2012-0004.
- [38] J. Pearl, From Bayesian Networks to Causal Networks, in: *Mathematical Models for Handling Partial Knowledge in Artificial Intelligence*, Springer US, Boston, MA, 1995, pp. 157–182. doi:10.1007/978-1-4899-1424-8{_}9.
- [39] J. Pearl, S. Russell, *Bayesian Networks*, Technical Report, 2003.
- [40] A. Zanga, E. Ozkirimli, F. Stella, A Survey on Causal Discovery: Theory and Practice, *International Journal of Approximate Reasoning* 151 (2022) 101–129. URL: <https://linkinghub.elsevier.com/retrieve/pii/S0888613X22001402>. doi:10.1016/j.ijar.2022.09.004.
- [41] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, *Causation, prediction, and search*, MIT press, 2000.
- [42] M. Scutari, Bayesian network models for incomplete and dynamic data, *Statistica Neerlandica* 74 (2020) 397–419. URL: <https://onlinelibrary.wiley.com/doi/10.1111/stan.12197>. doi:10.1111/stan.12197.
- [43] S. L. Lauritzen, The EM algorithm for graphical association models with missing data, *Computational Statistics and Data Analysis* 19 (1995). doi:10.1016/0167-9473(93)E0056-A.
- [44] N. Friedman, Learning Belief Networks in the Presence of Missing Values and Hidden Variables, in: *Proceedings of the Fourteenth International Conference on Machine Learning, ICML '97*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, pp. 125–133.
- [45] A. Zanga, A. Bernasconi, P. Lucas, H. Pijnenborg, C. Reijnen, M. Scutari, F. Stella, Risk Assessment of Lymph Node Metastases in Endometrial Cancer Patients: A Causal Approach, in: *Proceedings of the 1st Workshop on Artificial Intelligence For Healthcare (HC@AIxIA)*, volume 3307, 2022. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85145592167&partnerID=40>.
- [46] D. V. Nguyen, S.-W. Kim, Y.-T. Oh, O. K. Noh, Y. Jung, M. Chun, D. S. Yoon, Local recurrence in young women with breast cancer: Breast conserving therapy vs. mastectomy alone, *Cancers* 13 (2021) 2150. doi:10.3390/cancers13092150.
- [47] G. K. Jakubiak, K. Osadnik, M. Lejawa, S. Kasperczyk, T. Osadnik, N. Pawlas, Oxidative stress in association with metabolic health and obesity in young adults, *Oxidative Medicine and Cellular Longevity* 2021 (2021) 1–19. doi:10.1155/2021/9987352.

8. Supplementary Material

Table S1

Percentage of observed values for the variables in the model, categorized according to their role in the domain and stratified by cohort type (PBC, CBC).

Variable	Label	Category	PBC	CBC
Age>35 years	age35	Prognostic factor	100%	100%
Tumor Hystology	histology	Prognostic factor	100%	100%
Tumor grading	grade	Prognostic factor	0%	89%
Vascular invasion	vascular	Prognostic factor	0%	57%
Ki67 positivity	ki67	Prognostic factor	0%	92%
Tumor receptor status	receptors	Prognostic factor	0%	94%
Tumor dimension	pT	Prognostic factor	0%	65%
Lymph nodes involvement	pN	Prognostic factor	0%	68%
Death in 5 years	death_in_5y	Survival	90%	45%
Neo-adjuvant chemotherapy	chemo_neo	Treatment	100%	99%
Neo-adjuvant radiotherapy	radio_neo	Treatment	100%	99%
Neo-adjuvant target therapy	target_neo	Treatment	100%	99%
Neo-adjuvant hormon therapy	hormons_neo	Treatment	100%	99%
Surgery type	surgery	Treatment	100%	100%
adjuvant chemotherapy	chemo_adju	Treatment	100%	96%
adjuvant radiotherapy	radio_adju	Treatment	100%	96%
adjuvant target therapy	target_adju	Treatment	100%	96%
adjuvant hormon therapy	hormons_adju	Treatment	100%	96%
Dyslipidemia	dyslipidemia	CVDs risk factor	100%	5%
Hypertension	hypertension	CVDs risk factor	100%	3%
Type 2 Diabete Mellitus	t2db	CVDs risk factor	100%	1%
Cardiotoxicity	cardiotoxicity	CVDs risk factor	100%	0%
Ischemic heart disease	ischemic_heart_disease	CVDs risk factor	100%	0%
Cardiovascular disease in 5 years	cvds	Target variable	100%	0%

PBC = Population-based cohort; CBC = Clinical-based Cohort; CVDs = Cardiovascular diseases.